NTIRE 2019 Challenge on Video Deblurring: Methods and Results

Seungjun Nah Radu Timofte Sungyong Baik Seokil Hong Gyeongsik Moon
Sanghyun Son Kyoung Mu Lee Xintao Wang Kelvin C.K. Chan Ke Yu
Chao Dong Chen Change Loy Yuchen Fan Jiahui Yu Ding Liu
Thomas S. Huang Hyeonjun Sim Munchurl Kim Dongwon Park Jisoo Kim
Se Young Chun Muhammad Haris Greg Shakhnarovich Norimichi Ukita
Syed Waqas Zamir Aditya Arora Salman Khan Fahad Shahbaz Khan Ling Shao
Rahul Kumar Gupta Vishal Chudasama Heena Patel Kishor Upla Hongfei Fan
Guo Li Yumei Zhang Xiang Li Wenjie Zhang Qingwen He Kuldeep Purohit
A. N. Rajagopalan Jeonghun Kim Mohammad Tofighi Tiantong Guo
Vishal Monga

Abstract

This paper reviews the first NTIRE challenge on video deblurring (restoration of rich details and high frequency components from blurred video frames) with focus on the proposed solutions and results. A new REalistic and Diverse Scenes dataset (REDS) was employed. The challenge was divided into 2 tracks. Track 1 employed dynamic motion blurs while Track 2 had additional MPEG video compression artifacts. Each competition had 109 and 93 registered participants. Total 13 teams competed in the final testing phase. They gauge the state-of-the-art in video deblurring problem.

1. Introduction

Example-based video deblurring aims to recover the rich details and the sharp edges from blurry video frames based on a set of prior examples with blurry and sharp videos. The loss of sharpness can be caused by various sources, typically by the motions during the exposure. Hand-held cameras are prone to shake while the multiple objects in the scene can act with translation, rotation, deformation, etc. The presence of blur in the videos makes it hard to recognize textures in the scene rendering the videos visually unpleasing. Deblurring is generally an ill-posed problem since infinitely many latent sharp frames are in the large solution space for a single blurry frame.

In the field of image/video restoration, deblurring has received much attention and many approaches have been proposed. However, evaluating the accuracy of a deblurring algorithm is a tricky issue as it is hard to acquire a pixel-level aligned pair of blurry and sharp image simultaneously. Köhler et al. [10] used a robot to record camera motions and replayed it to capture the same scenes with different exposures. They successfully modeled camera shake blurs for 4 different scenes and 12 different kernels. More recently, new techniques were proposed to synthesize realistic blurry images from high-speed cameras [13, 18] and high-resolution videos [27]. By averaging high-frame-rate (240 fps) video frames to generate a blurry frame, realistic blurs of dynamic scenes could be generated.

The proposed datasets enabled quantitative evaluation of the proposed deblurring methods and inspired the training of various deep neural networks [13, 18, 27, 8, 20, 9]. However, the problem of modeling blurs with realism is not solved yet. 240 fps is still too slow to capture the fast motion of objects [27] and simple averaging of frames does not accurately model the nonlinear property of the camera imaging pipeline [13]. Also, the typical lossy compression of video which makes deblurring difficult, is not considered. Furthermore, the high-frame-rate video recordings usually have less number of effective pixels than the frame resolution. This is because the camera sensors cannot be fully accessed by the processors during the short readout time in a duty cycle. It makes the video reference frame quality to be lower than typical frame-rate videos or static photographs.

The NTIRE 2019 video deblurring challenge is a step forward in benchmarking and training of video deblurring algorithms. It uses REalistic and Dynamic Scenes (REDS)

S. Nah (seungjun.nah@gmail.com, Seoul National University), R. Timofte, S. Baik, S. Hong, G. Moon, S. Son and K. M. Lee are the NTIRE 2019 challenge organizers, while the other authors participated in the challenge.
Appendix A contains the authors' teams and affiliations.
NTIRE webpage: http://www.vision.ee.ethz.ch/ntire19/

(a) Sharp Uncompressed (b) Blurry (c) Blurry + Compressed

Figure 1: Visualization of the REDS dataset used for NTIRE 2019 Video Deblurring Challenge. The degradations include dynamic motion blur and lossy video compression artifacts.

dataset [12] consisting of 30000 reference frames with two types of degradation: dynamic motion blurs and MPEG video compression artifacts. The REDS dataset is introduced in [12] along with a study of challenge results. In the next, we describe the NTIRE 2019 video deblurring challenge, present and discuss the results and describe the methods.

2. The Challenge

The objectives of the NTIRE 2019 challenge on video deblurring are: (i) to gauge and push forward the state-of-the-art in video deblurring; (ii) to compare different solutions; (iii) to promote REDS, a novel large-scale high-quality video dataset; and (iv) to promote more challenging video deblurring settings.

2.1. REDS Dataset

As a step forward from the previously deblurring datasets, a novel dataset is promoted, namely REDS dataset [12]. It consists of 300 video sequences having length of 100 frames with 720×1280 resolution. 240 sequences are for training, 30 for validation and the rest 30 for testing purposes. The frames are of high quality in terms of the reference frames, diverse scenes and locations, realistic approximation of motion blurs, and the used standard lossy compression artifacts. REDS covers a large diversity of contents, people, handmade objects and environments (cities).

All the videos used to create the dataset are manually recorded with GoPro HERO6 Black. They were originally recorded in 1080×1920 resolution at 120 fps. We calibrated the camera response function using [17] with regularization. Then the frames are interpolated [15] to virtually increase the frame rate up to 1920 fps so that averaged frames could exhibit smooth and realistic blurs without spikes and step artifacts. Then, the virtual frames are averaged in the signal space to mimic camera imaging pipeline [13]. To suppress noise and compression artifacts in the reference frames and to increase the effective number of pixels per resolution, the synthesized blur and the corresponding sharp frames are downscaled to 720×1280 resolution. The resulting blurry video frames simulate 24 fps videos captured at duty cycle $\tau = 0.8$.

2.2. Tracks and competitions

Track 1: Clean facilitates a deployment of recently proposed methods for the task of example-based video deblurring. It assumes that the quality degradation is caused by the camera shakes and the motion of objects that are the main assumptions in dynamic scene deblurring. Each blurry frame is obtained by the procedure described in Section 2.1 and with more details in [12].

Track 2: Compression artifacts goes one step ahead and considers the actual compression of video recordings. The blurry frames in a sequence that are saved independently after synthesis are collected to be saved as a single mp4 (MPEG-4 Part 14) video clip. We used MATLAB to save the video at 60% quality. Note that the degradation degree of each frame due to compression may not uniform.

Competitions Both video deblurring challenge tracks are hosted as Codalab competitions. CodaLab platform was used for all of the NTIRE 2019 challenge competitions. Each participant is required to register to the CodaLab challenge tracks to access the data and submit their deblurred results.

Challenge phases (1) Development (training) phase: the participants got both the blurry, the additionally compressed and the sharp train video frames and the blurry frames of the

validation set. The participants had the opportunity to test their solutions on the blurry validation frames and to receive feedback by uploading their results to the server. Due to the large-scale of the validation dataset, every 10^{th} frame was involved in evaluation. A validation leaderboard is available; (2) Final evaluation (test) phase: the participants got the sharp reference validation frames with the blurry test frames. They had to submit both the deblurred frames and a description of their methods and code/executables before the challenge deadline. One week later, the final results were made available to the participants. The final results reflect the performance on every frame of the test set.

Evaluation protocol The Peak Signal-to-Noise Ratio (PSNR) measured in deciBels (dB) and the Structural Similarity index (SSIM) [26] computed between a result frame and the ground truth are the quantitative measures. The higher the scores are the better the restoration fidelity to the ground truth frame. A rim of 1 pixel is ignored in the evaluation.

3. Challenge Results

From 109 and 93 registered participants for the competitions, 13 teams entered in the final phase and submitted results, codes/excutables, and factsheets. Table 1 reports the final scoring results of the challenge and Table 2 shows the runtimes and the major details for each entry. Section 4 describes briefly the method of each team while in the Appendix A are the team members and their affiliations.

Use of temporal information All the proposed methods use the end-to-end deep learning and employ the GPU(s) for both training and testing. Interestingly, in contrast to the recent RNN-based video deblurring methods [8, 27, 9], most teams (HelloVSR, UIUC-IFP, KAIST-VICLAB, BMIPL_UNIST_DJ, IPCV_IITM, JeonghunKim, iPAL-Deblur) aggregated several video frames in channel dimension and let CNN learn the temporal relation to deblur a target frame. TTI used a recurrent model inspred from DBPN [4]. TeamInception, Game of tensors, and KSC proposed single image deblurring methods.

Restoration fidelity HelloVSR, UIUC-IFP, and KAIST-VICLAB are the best scoring teams. HelloVSR is ***the winner of NTIRE 2019 Video Deblurring Challenge***. HelloVSR achieves 36.96 dB for Track 1 and 31.69 dB for Track 2 that are $+10.83$ dB and $+6.29$ dB better than the input blurry video, respectively. HelloVSR team achieves the best results for both of the competition tracks. Their solution shows significant improvements compared to the other submitted solutions and is also consistent in NTIRE 2019 Video Super-Resolution Challenge [14].

Runtime / efficiency In Fig. 2 and 3, we plot the running time per image vs. achieved PSNR performance for both tracks. Interestingly, HelloVSR team showed good efficiency in running time, compared to other high ranked solutions. KSC team showed good trade off between runtime and quality of the results. In takes 0.78 s per frame for Track 1, on Titan X Pascal GPU while most other solutions require more than 1 second per frame.

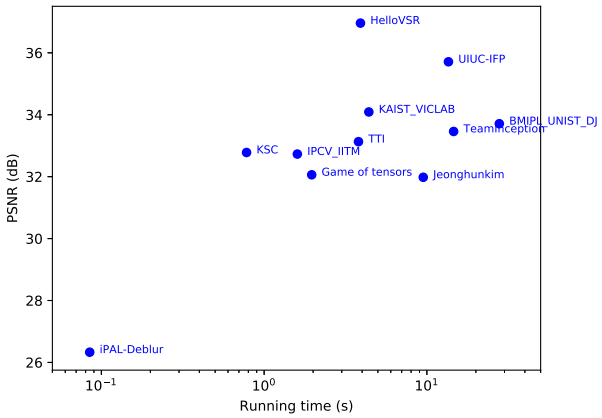

Figure 2: Runtime vs. performance for Track 1: Clean.

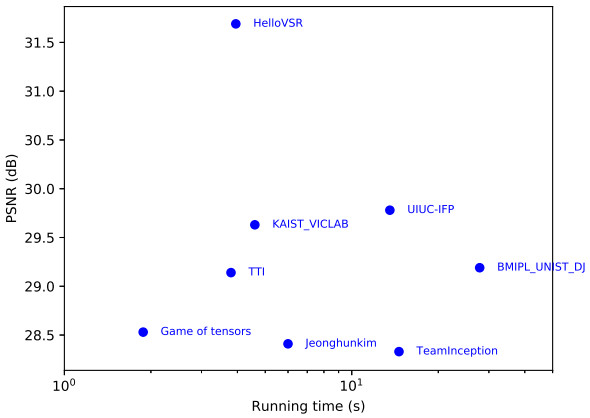

Figure 3: Runtime vs. performance for Track 2: Compression artifacts.

Ensembles Many solutions used self-ensemble [22] that averages the results from flipped and rotated inputs at test time. HelloVSR did not use rotation to reduce computation.

Train data REDS dataset [12] has 24000 train frames and all the participants found the amount of data to be sufficient for training their models. Training data augmentation strategy [22] such as flips and rotations by 90 degrees were employed by most of the participants.

Conclusions From the analysis of the presented results, we conclude that the proposed methods gauge the state-of-the-art performance in video deblurring. The methods proposed by the high ranking teams (HelloVSR, UIUC-IFP, KAIST-VICLAB) exhibit consistent superiority in both tracks in terms of PSNR and SSIM.

Team	Author	Track 1: Clean PSNR	Track 1: Clean SSIM	Track 2: Compression artifacts PSNR	Track 2: Compression artifacts SSIM
HelloVSR	xixihaha	36.96 (1)	0.9657	31.69 (1)	0.8783
UIUC-IFP	fyc0624	35.71 (2)	0.9522	29.78 (2)	0.8285
KAIST-VICLAB	KAIST_VICLAB	34.09 (3)	0.9361	29.63 (3)	0.8261
BMIPL_UNIST_DJ	BMIPL_UNIST_JS	33.71 (4)	0.9363	29.19 (4)	0.8190
TTI	iim_lab	33.13 (6)	0.9198	29.14 (5)	0.8145
TeamInception	swz30	33.46 (5)	0.9293	28.33 (9)	0.7976
Game of tensors	rkgupta	32.06 (9)	0.9070	28.53 (6)	0.8034
KSC	fanhongfei	32.78 (7)	0.9187	-	-
IPCV_IITM	kuldeeppurohit3	32.73 (8)	0.9147	-	-
Jeonghunkim	JeonghunKim	31.98(10)	0.9061	28.41 (8)	0.7962
Game of tensors	rahul12122	31.76(11)	0.9025	28.44 (7)	0.8014
withdrawn team		29.56(12)	0.8474	27.76(10)	0.7801
iPAL-Deblur	mo.tofighi	26.33(13)	0.7491	-	-
no processing	*baseline*	26.13	0.7749	25.40	0.7336

Table 1: NTIRE 2019 Video Deblurring Challenge results on the REDS test data. HelloVSR team is the challenge winner with UIUC-IFP and KAIST-VICLAB coming on 2^{nd} and 3^{rd} place, respectively, with consistent performance on both tracks.

Team	Track 1 Clean	Track 2 Compression artifacts	Platform	GPU (at runtime)	Ensemble / Fusion (at runtime)
HelloVSR	3.908	3.950	PyTorch	TITAN Xp	flip (x4)
UIUC-IFP	13.570	13.570	PyTorch	Tesla V100	flip/rotation (x8)
KAIST-VICLAB	4.540	4.540	TensorFlow	TITAN Xp	flip/rotation (x8)
BMIPL_UNIST_DJ	27.870	27.870	PyTorch	TITAN V	flip/rotation (x8)
TTI	3.800	3.800	PyTorch	TITAN X	-
TeamInception	14.600	14.600	PyTorch	Tesla V100	flip/rotation (x8)
Game of tensors	1.960	1.880	TensorFlow	TITAN X	flip/rotation (x8)
KSC	0.780	-	TensorFlow	TITAN X	-
IPCV_IITM	1.600	-	PyTorch	TITAN X	flip/rotation (x8)
Jeonghunkim	9.500	6.000	TensorFlow	TITAN X	flip/rotation (x8)
Game of tensors	0.524	0.500	TensorFlow	TITAN X	-
withdrawn team	398.000	398.000	-	-	-
iPAL-Deblur	0.085	-	PyTorch	TITAN Xp	-

Table 2: Reported runtimes per frame on REDS test data and details from the factsheets.

4. Challenge Methods and Teams

4.1. HelloVSR team

HelloVSR team proposes the EDVR framework [23], which takes $2N+1$ degraded frames as inputs and generates one restored frame, as shown in Fig. 4. First, the high-resolution inputs are first downsampled by a factor of 4. Then most computation is done in the low-resolution space, which largely saves the computational cost. To alleviate the effects of blurry frames on alignment, a PreDeblur module is used to pre-process the blurry inputs before alignment. Each neighboring frame is aligned to the reference frame by the PCD alignment module at the feature level. The TSA fusion module is used to effectively fuse the aligned features. The fused features then pass through a reconstruction module, which consists of several residual blocks [11] in EDVR and can be easily replaced by any other advanced modules in image restoration [13, 31, 30, 25]. The upsampling operation is performed at the end of the network to resize the features back to the original input resolution. Finally, the restored frame is obtained by adding the predicted image residual to the input reference frame [7]. Note that EDVR is a generic architecture also suitable for other video restoration tasks, such as super-resolution.

To address large and complex motions between frames, which are common in the REDS dataset, they propose a

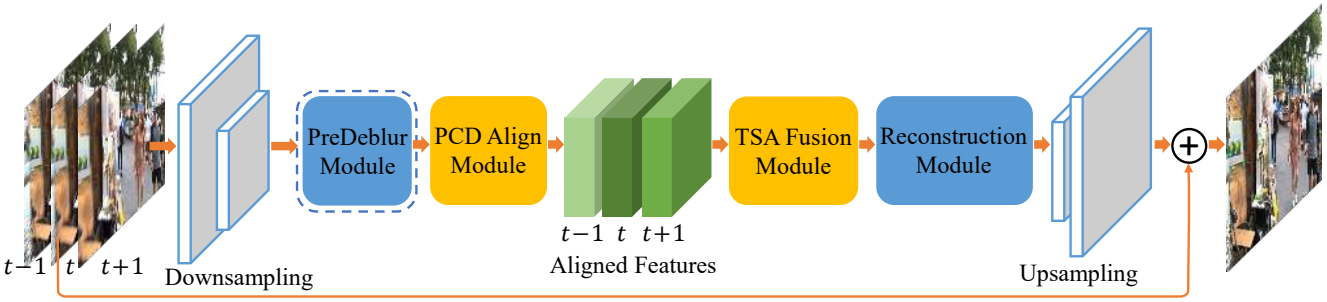

Figure 4: HelloVSR team: the proposed EDVR framework.

Pyramid, Cascading and Deformable convolution (PCD) alignment module. In this module, deformable convolutions [2, 21] is adopted to align frames at the feature level. They use a pyramid structure that first aligns features in lower scales with coarse estimations, and then propagates the offsets and aligned features to higher scales to facilitate precise motion compensation, which is similar to the notion adopted in optical flow estimation [6, 19]. Moreover, an additional deformable convolution is cascaded after the pyramidal alignment to further improve the robustness of alignment. The overview of PCD module is shown in Fig. 5.

Since different frames and locations are not equally informative due to the imperfect alignment and imbalanced blur among frames, a Temporal and Spatial Attention (TSA) fusion module is designed to dynamically aggregate neighboring frames in pixel-level, as shown in Fig. 5. Temporal attention is introduced by computing the element-wise correlation between the reference frame and each neighboring frame in an embedding space. The correlation coefficients then weigh each neighboring feature at each location. The weighted features from all frames are then convolved and fused together. After the fusion, they further apply spatial attention [28, 24, 30] to assign weights to each location in each channel to exploit cross-channel and spatial information more effectively.

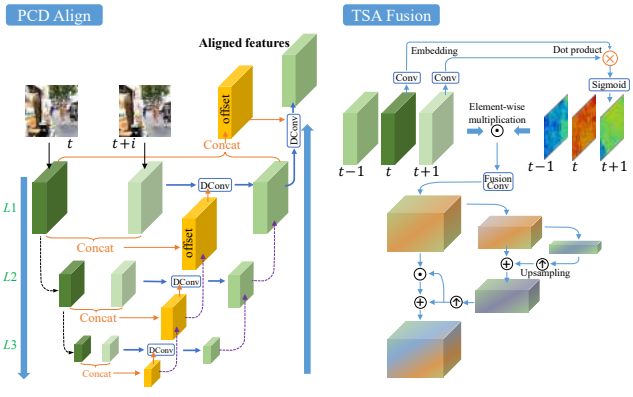

Figure 5: PCD alignment module and TSA fusion module in EDVR.

They also use a two-stage strategy to further boost the performance. Specifically, a similar but shallower EDVR network is cascaded to refine the output frames of the first stage. The cascaded network can further remove the severe motion blur that cannot be handled by the preceding model and alleviate the inconsistency among output frames.

4.2. UIUC-IFP team

UIUC-IFP team proposes a new method, WDVR, which is based on WDSR [29, 3]. To achieve a better speed-accuracy trade-off, they investigate the intersection of three dimensions in deep video restoration networks: spatial, channel, and temporal. They enumerate various network architectures ranging from 2D convolutional models to 3D convolutional models and delve into their gains and losses in terms of training time, model size, boundary effects, prediction accuracy, and the visual quality of the restored videos. Under a strictly controlled computational budget, they explore the designs of each residual building block in a video restoration network, which consists of a mixture of 2D and 3D convolutional layers.

4.3. KAIST-VICLAB team

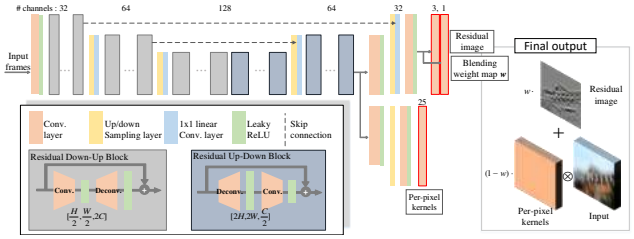

Figure 6: KAIST-VICLAB team: overall pipeline of the model.

KAIST-VICLAB team proposes a convolutional neural network to predict motion deblur kernel for each pixel location. Each 2-dimensional deblur kernel is applied to the corresponding input blurry pixel and its square neighbor pixels to generate a clean pixel. Output pixel values are borrowed from the surrounding values, as the output pixel is a dot product of its neighboring pixels and their corresponding

coefficients, the deblur kernel. Therefore, this *local* convolution estimates coarse (low-frequency) output. At the same time, the network outputs RGB residual image which is added to the above locally convolved output. Since the image created by local convolution is responsible for low-frequency component of the clean image, the other branch of the network only needs to produce a residual image, which is simpler than hallucinating the component of the pixel values from scratch. As a result, the RGB output image of the convolutional neural network can concentrate on the details (high-frequency) of the scene. The locally convolved output and the residual image are linearly combined to be the final output with a weight map which is another branch of the network. Hence, there are three outputs of the proposed network. The overall architecture is depicted in Fig. 6. They also propose a spatial-bottleneck block as a base building block of their network. The proposed model consists of mainly two parts, encoder and decoder. For the encoder part, residual encoder blocks are the base blocks. Input tensor is fed into a convolutional layer with stride 2 and a transposed convolution with stride 2 in the residual encoder block. The first convolutional layer halves the spatial size while doubling the number of channels of the input tensor. The following transposed layer reversely doubles the spatial size and reduce the number of channels to the original size. At the end of the block, input tensor is added for the purpose of residual learning. This procedure enlarges the receptive field size of the network while retaining details by residual learning. For the decoder part, residual decoder blocks are the base building blocks. The residual decoder blocks are similar to the residual encoder blocks, but the order of convolutional and transposed convolutional layer is reversed.

4.4. BMIPL_UNIST_DJ team

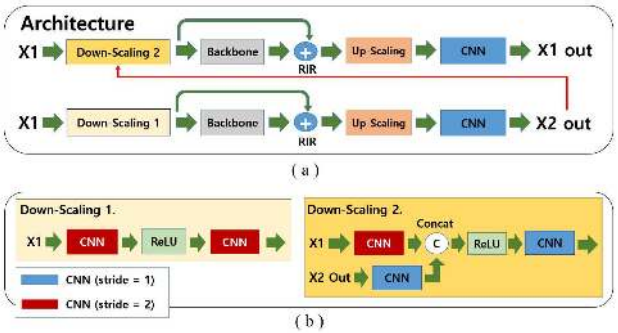

Figure 7: BMIPL_UNIST_DJ team: proposed pipeline.

BMIPL_UNIST_DJ team proposes a network [16] with a new module base, as shown in Fig. 7. The network first down-scales the input color image with the size of $W \times H \times 3$ to the feature maps with the size of $W/4 \times H/4 \times 64$, using the "Down-Scaling 1" module in Fig. 7(b). Then, these feature maps are fed into the backbone network as well as residual in residual (RIR) skip connection to yield initial deblurred feature maps in the spatial dimension of $W/4 \times H/4$. Then, up-scaling and CNN will result in the intermediate deblurred image estimate with the size of $W/2 \times H/2 \times 3$. This result is combined with the down-scaled blurred image in the "Down-Scaling 2" module in Fig. 7 (b) for the deblurring at the scale of $W/2 \times H/2$ using another backbone network, RIR, and up-scaling process to yield the final deblurred output image with the size of $W \times H$. And they use residual channel attention network (RCAN) [30] as a backbone of their network. They set the number of res-group as 10 and the number of res-block as 20.

4.5. TTI team

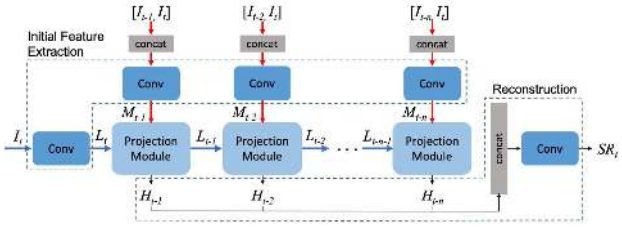

Figure 8: TTI team: proposed scheme.

TTI team proposes a network architecture in Fig. 8 inspired by a single image super-resolution model [4]. They integrate spatial and temporal contexts from the consecutive video frames using a recurrent encoder-decoder module. The module fuses multi-frame information with more traditional single image deblurring computation path for the target frame. In contrast to most of the prior works where frames are pooled together by stacking or warping, their model, Recurrent Back-Projection Network (RBPN), treats each context frame as a separate source of information. These sources are combined in an iterative refinement framework inspired by the idea of back-projection. This is aided by explicitly representing estimated inter-frame motion with respect to the target frame, rather than explicitly aligning frames.

4.6. TeamInception team

TeamInception team proposes a deep residual network with spatial and depth-wise attention to address the problem of video deblurring. The complete framework is shown in Fig. 9. Inspired from the work of [30] on super-resolution, they propose a method that is recursive in nature. The main idea is to gradually remove the effect of blurring from the input frame and recover the sharp image. At the entry point of the RDAN (recursive dual attention network), they employ a convolutional layer that takes as input the blurred image and extracts low-level features. The fea-

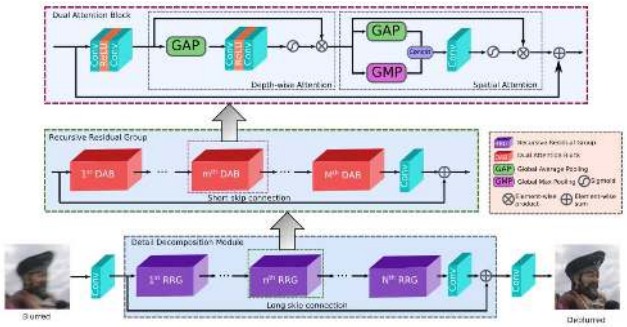

Figure 9: TeamInception team: proposed solution.

ture maps become input to the detail decomposition module (DDM), which contains N number of recursive residual groups (RRGs). The goal of RRG is to progressively recover the information related to the desired sharp image from the input blurred image. Each RRG further employs M number of dual attention blocks (DABs). Features that are less important get suppressed in the DAB, and only useful information is propagated onward. To discern such features, they apply two types of attention mechanisms in DAB depth-wise attention, and spatial attention.

Depth-wise attention branch models the relationships among channels and adaptively rescales the feature responses. It does so by first applying a global average pooling (GAP) operation on the input tensor of size $H' \times W' \times C'$ to obtain a global descriptor of size $1 \times C'$. Next, a convolutional layer is applied to the global descriptor that reduces its channel dimension to $1 \times \frac{C'}{r}$, followed by ReLU non-linearity. The downscaling factor r is set to 16 in their experiments. The descriptor then passes through a channel-upscaling convolutional layer that yields a descriptor of size $1 \times C'$, which they activate with a sigmoid gating function. Finally, the descriptor enriched with channel statistics is used to rescale the input tensor. Spatial attention branch receives the input feature tensor from the depth-wise attention branch. At the beginning of the spatial attention branch, two operations namely global average pooling (GAP) and global max pooling (GMP) are independently applied across channels, resulting in 2 single channel feature maps that are concatenated together before being passed to the convolutional layer. The output of the convolutional layer is activated by the sigmoid function, thus generating an activation map with aggregated responses. This activation map is finally multiplied, element-wise, with the input tensor. In each DAB, the features belonging to the latent deblurred image are bypassed via skip connections while other features go through subsequent DABs for further detail decomposition. The last convolutional layer of our network receives deep features from the last RRG module (after convolution) and yields the final image having the same resolution as of the input image.

In order to minimize the distance between networks output y and the ground truth image $\hat{y}$, they use a multi-criterion loss function in their RDAN network.

$$\mathcal{L}_f = \alpha \mathcal{L}_1(\hat{y}, y) + \beta \mathcal{L}_{MS-SSIM}(\hat{y}, y) + \gamma \mathcal{L}_{VGG}(\hat{y}, y) \tag{1}$$

The first term ($\mathcal{L}_1$ loss) and the second term (multi-scale structural similarity measure) compute differences between the network's output and the ground truth directly at the pixel-level, whereas the last term of the loss function compares the deep feature representations of the output and ground-truth images extracted with the VGG network pre-trained on the ImageNet dataset.

$$\mathcal{L}_{VGG}(\hat{y}, y) = \frac{1}{N} \|\phi(\hat{y}) - \phi(y)\|_2^2 \tag{2}$$

where N denotes the total number of pixels in the image. In their experiments, they use `conv2` layer after ReLU of the VGG-16 network.

4.7. Game of tensors team

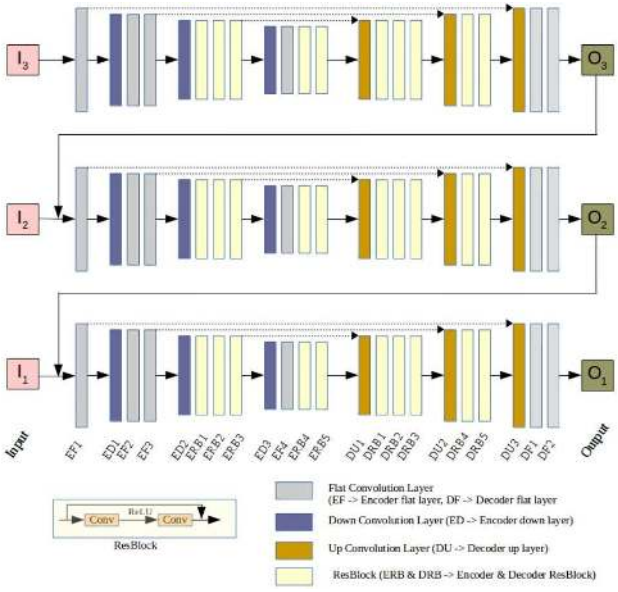

Figure 10: Game of tensors team: overall pipeline of the model.

Game of tensors team proposes a auto-encoder-based multi-scale deep neural network shown in Fig. 10.

4.8. KSC team

KSC team proposes a novel network structure called two-stage scale-recurrent network. The network contains two stages: pre-trained SRN-DeblurNet(SRN) [20] and the second network. Pre-trained network can successfully generate deblurred results. The second network tries to solve harder work by deepening different scale networks which

integrated with the multi scale outputs of pre-trained network. For better results, they fine tune SRN-DeblurNet by abandoning LSTM, enhancing data, changing upsampling method to bicubic, filtering data according to PSNR and gradient, using adaptive learning rate, and four scales during inference. And finally, they use Extreme Channel Prior Embedded Network(ECPE-Net) [1] as the second network.

4.9. IPCV_IITM team

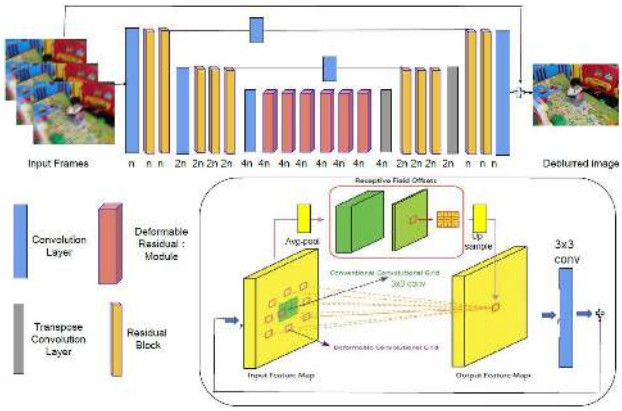

Figure 11: IPCV_IITM team: proposed model.

IPCV_IITM team proposes a residual encoder-decoder structure-based network. The encoder sub-network extracts features from input frames and reduces the spatial size by a factor of 4, on which the spatially adaptive modules of the network operate. The spatially adaptive modules allow image-dependent filter offset estimation within residual blocks and are present in the middle of the network (referred to as Deformable Residual Modules). The overall architecture is shown in Fig. 11, where n represents the number of channels in the first feature map.

For Deformable Residual Modules, they adopt DCN [2], which enables local transformation learning in an efficient manner. While maintaining filter weights invariant to the input, the deformable convolution first learns a dense offset map from the input and then applies it to the regular feature map for re-sampling.

In their final network, 6 DRBs are present in the mid-level layer of the network, where the spatial resolution of features is the smallest and hence offset estimation causes marginal increase in processing time. Additive link between encoder and decoder feature-maps grant the benefits of reusing common features with low redundancy. The network contains very few parameters (< 3 Million).

4.10. JeonghunKim team

JeonghunKim team proposes *Deblur network considering detail intensively (PDNet : Pyramid and Detail Network)*. In general, only using pyramid has the weak detail despite large receptive field. So they propose context refine network following the last pyramid level and Detail loss and Correlation loss considering detail intensively. These components help network learn detail information strongly. As the input of context refine network, they compute the 4 gradient of last pyramid level output image with respect to up, down, left, right direction. By this strategy, network can easily refer neighbor information from all directions. And Detail loss is based on detail part of network output image which is computed by subtracting the guided filtering [5], which is part of low frequency, from network output image. Correlation loss is based on the inverse correlation between network output image and ground truth. Using $\mathcal{L}_2$ loss only, there is obviously limit on deblurring image because $\mathcal{L}_2$ loss just averages out the difference between two image intensities that results in the vague boundary of blurred object.

4.11. iPAL-Deblur team

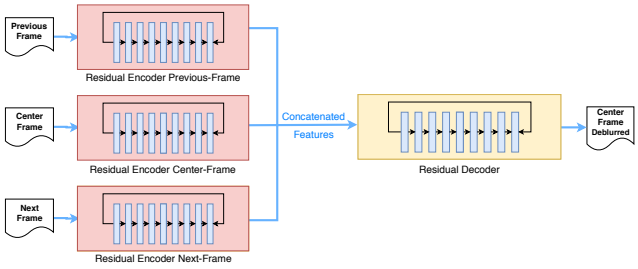

Figure 12: iPAL-Deblur team: proposed solution.

iPAL-Deblur team proposes an efficient network structure that makes use of neighboring frames to learn both spatial (in-frame) and temporal (neighboring frames) features in training of the network. They call the proposal Spatio-Temporal features for Deep Video Deblurring (STDVD). The input to STDVD network is each frame of the video along with its neighboring frames (one or more frames around each frame). In a departure from most existing deep video deblurring methods, they design a custom encoder for each frame. The output features of these encoders are then concatenated and fed into a decoder that generates the deblurred frame. The STDVD framework is illustrated in Fig. 12. The idea behind using multiple frames for deblurring a single video frame is that while recording real world videos, adjacent frames may have different blur artifacts. STDVD can train for pixels of the center frame using the possibly sharp pixels from the neighboring ones. They show that STDVD is significantly faster owing to their network structure and allows for real-time processing. As an example, STDVD generates the deblurred frame of 1280×720 pixel resolution in almost 0.03 seconds (33 fps) on a typical NVIDIA GPU which is sufficient for processing a video at 24 fps.

Acknowledgments

We thank the NTIRE 2019 sponsors: OPPO Mobile Corp., Ltd., NVIDIA Corp., HUAWEI Technologies Co. Ltd., SAMSUNG Electronics Co., Ltd., Amazon.com, Inc., MediaTek Inc., and ETH Zurich.

A. Teams and affiliations

NTIRE2019 team

Title: NTIRE 2019 Challenge on Video Deblurring
Members: *Seungjun Nah*[1] *(seungjun.nah@gmail.com)*, Radu Timofte[2], Sungyong Baik[1], Seokil Hong[1], Gyeongsik Moon[1], Sanghyun Son[1], Kyoung Mu Lee[1]
Affiliations:
[1] Department of ECE, ASRI, SNU, Korea
[2] Computer Vision Lab, ETH Zurich, Switzerland

HelloVSR team

Title: EDVR: Video Restoration with Enhanced Deformable Convolutional Networks
Members: *Xintao Wang*[1] *(xintao.alpha@gmail.com)*, Kelvin C.K. Chan[2], Ke Yu[1], Chao Dong[3], Chen Change Loy[2]
Affiliations:
[1] The Chinese University of Hong Kong, Hog Kong
[2] Nanyang Tenchnological University, Singapore
[3] Shenzhen Institutes of Advanced Tenchnology, Chinese Academy of Sciences, China

UIUC-IFP team

Title: Wide Activation for Efficient and Accurate Video Deblurring
Members: *Yuchen Fan (yuchenf4@illinois.edu)*, Jiahui Yu, Ding Liu, Thomas S. Huang
Affiliations:
University of Illinois at Urbana-Champaign, US

KAIST-VICLAB team

Title: A Deep Motion Deblurring Network based on Per-Pixel Adaptive Kernels with Residual Down-Up and Up-Down Modules
Members: *Hyeonjun Sim (flhy5836@kaist.ac.kr)*, Munchurl Kim
Affiliations:
Korea Advanced Institute of Science and Technology (KAIST), Korea

BMIPL_UNIST_DJ team

Title: Down-Scaling with Learned Kernels in Multi-Scale Deep Neural Networks for Non-Uniform Single Image Deblurring
Members: *Dongwon Park (dong1@unist.ac.kr)*, Jisoo Kim, Se Young Chun
Affiliations:
Ulsan National Institute of Science and Technology (UNIST), Korea

TTI team

Title: Recurrent Back-Projection Networks
Members: *Muhammad Haris*[1] *(mharis@toyota-ti.ac.jp)*, Greg Shakhnarovich[2], Norimichi Ukita[1]
Affiliations:
[1] Toyota Technological Institute (TTI), Japan
[2] Toyota Technological Institute at Chicago (TTIC), US

TeamInception team

Title: Video Deblurring via Recursive Residual Network with Dual Attention
Members: *Syed Waqas Zamir (waqas.zamir@inceptioniai.org)*, Aditya Arora, Salman Khan, Fahad Shahbaz Khan, Ling Shao
Affiliations:
Inception Institute of Artificial Intelligence (IIAI), UAE

Game of tensors team

Title: Game of tensors
Members: *Rahul Kumar Gupta (singhalrahul222@gmail.com)*, Vishal Chudasama, Heena Patel, Kishor Upla
Affiliations:
Sardar Vallabhbhai National Institute of Technology, India

KSC team

Title: Two-stage scale-recurrent network
Members: *Hongfei Fan (fanhongfei@kingsoft.com)*, Guo Li, Yumei Zhang, Xiang Li, Wenjie Zhang, Qingwen He
Affiliations:
Kingsoft Cloud, China

IPCV_IITM team

Title: Dynamic Residual Network for Efficient Video Deblurring
Members: *Kuldeep Purohit*

(kuldeeppurohit3@gmail.com), A. N. Rajagopalan
Affiliations:
Indian Institute of Technology Madras, India

JeonghunKim team

Title: Deblur network considering detail intensively
Members: *Jeonghun Kim (yblues80@kaist.ac.kr)*
Affiliations:
Korea Advanced Institute of Science and Technology (KAIST), Korea
Samsung Electronics, Korea

iPAL-Deblur team

Title: Deep Networks for Video Deblurring using Spatiotemporal Information
Members: *Mohammad Tofighi (mqt5352@psu.edu)*, Tiantong Guo, Vishal Monga
Affiliations:
Electrical Engineering Department, Pennsylvania State University, US

References

[1] Jianrui Cai, Wangmeng Zuo, and Lei Zhang. Extreme channel prior embedded network for dynamic scene deblurring. *arXiv preprint arXiv:1903.00763*, 2019. 8

[2] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2017. 5, 8

[3] Yuchen Fan, Jiahui Yu, and Thomas S. Huang. Wide-activated deep residual networks based restoration for bpg-compressed images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 5

[4] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3, 6

[5] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. In *The European Conference on Computer Vision (ECCV)*, September 2010. 8

[6] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. LiteFlowNet: A lightweight convolutional neural network for optical flow estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 5

[7] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 4

[8] Tae Hyun Kim, Kyoung Mu Lee, Bernhard Scholkopf, and Michael Hirsch. Online video deblurring via dynamic temporal blending network. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2017. 1, 3

[9] Tae Hyun Kim, Mehdi S. M. Sajjadi, Michael Hirsch, and Bernhard Scholkopf. Spatio-temporal transformer network for video restoration. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1, 3

[10] Rolf Köhler, Michael Hirsch, Betty Mohler, Bernhard Schölkopf, and Stefan Harmeling. Recording and playback of camera shake: Benchmarking blind deconvolution with a real-world database. In *The European Conference on Computer Vision (ECCV)*, October 2012. 1

[11] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 4

[12] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenges on video deblurring and super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2, 3

[13] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2, 4

[14] Seungjun Nah, Radu Timofte, Shuhang Gu, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Kyoung Mu Lee, Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, Chen Change Loy, Yuchen Fan, Jiahui Yu, Ding Liu, Thomas S. Huang, Xiao Liu, Chao Li, Dongliang He, Yukang Ding, Shilei Wen, Fatih Porikli, Ratheesh Kalarot, Muhammad Haris, Greg Shakhnarovich, Norimichi Ukita, Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, Jiayi Ma, Hang Dong, Xinyi Zhang, Zhe Hu, Kwanyoung Kim, Dong Un Kang, Se Young Chun, Kuldeep Purohit, A. N. Rajagopalan, Yapeng Tian, Yulun Zhang, Yun Fu, Chenliang Xu, A. Murat Tekalp, M. Akin Yilmaz, Cansu Korkmaz, Manoj Sharma, Megh Makwana, Anuj Badhwar, Ajay Pratap Singh, Avinash Upadhyay, Rudrabha Mukhopadhyay, Ankit Shukla, Dheeraj Khanna, A.S. Mandal, Santanu Chaudhury, Si Miao, Yongxin Zhu, and Xiao Huo. Ntire 2019 challenge on video super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 3

[15] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[16] Dongwon Park, Jisoo Kim, and Se Young Chun. Downscaling with learned kernels in multi-scale deep neural networks for non-uniform single image deblurring. *arXiv preprint arXiv:1903.10157*, 2019. 6

[17] Mark A. Robertson, Sean Borman, and Robert L. Stevenson. Dynamic range improvement through multiple exposures. In *The IEEE International Conference on Image Processing (ICIP)*, October 1999. 2

[18] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[19] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 5

[20] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 7

[21] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. TDAN: Temporally deformable alignment network for video super-resolution. *CoRR*, abs/1812.02898, 2018. 5

[22] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[23] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 4

[24] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 5

[25] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision (ECCV) Workshops*, September 2018. 4

[26] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *The IEEE Transactions on Image Processing (TIP)*, 13(4):600–612, 2004. 3

[27] Patrick Wieschollek, Michael Hirsch, Bernhard Scholkopf, and Hendrik P. A. Lensch. Learning blind motion deblurring. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2017. 1, 3

[28] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. In *The European Conference on Computer Vision (ECCV)*, September 2018. 5

[29] Jiahui Yu, Yuchen Fan, Jianchao Yang, Ning Xu, Zhaowen Wang, Xinchao Wang, and Thomas S. Huang. Wide activation for efficient and accurate image super-resolution. *CoRR*, abs/1808.08718, 2018. 5

[30] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *The European Conference on Computer Vision (ECCV)*, September 2018. 4, 5, 6

[31] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 4